Revista Colombiana de Estadística
Special Issue: Colombian and Brazil Graduate Programs
December 2025, volume 48, issue 3, pp. 553–571
http://doi.org/10.15446/rce.v48n3.123501

Promotion Time Cure Regression Model with Power Piecewise Exponential Distribution for Credit Scoring Data

Modelo de regresión con cura de tiempo de promoción con distribución potencia exponencial por partes para datos de calificación crediticia

Leonardo de Miranda Pinheiro[1,a], Silvana Schneider[2,b], Paulo Cerqueira dos Santos Junior[3,c]

[1]Graduate Program of Statistics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

[2]Graduate Program of Statistics, Department of Statistics, Federal University of Rio Grande do Sul, Porto Alegre, Brazil

[3]Graduate Program in Mathematics and Statistics, Department of Statistics, Federal University of Pará, Pará, Brazil

Abstract

In the financial context, survival analysis has been used in a variety of ways, like in the study (Borelli & Lucena, 2022), which aims to estimate the time until the recovery of overdue credit portfolios, with a focus on pricing non-performing credit portfolios. As in the study by Ramirez (2016), which aimed to estimate the time until default, evaluating the conditioning variables for credit delay. One of the objectives of this paper is to analyze a real dataset about credit, to model the time until the customer of a financial company, located in Rio Grande do Sul - Brazil, becomes defaulter. Therefore, in this paper it is proposed a model for survival data with cure rate, in which the distribution of the times is adjusted by the power piecewise exponential (PPE) distribution. The other objective is to propose a model that has not yet been explored jointly in the literature, through the construction of a long-term survival model, considering the approach of the promotion time models (Yakovlev & Tsodikov, 1996) and the power piecewise exponential distribution (Gómez et al., 2017). To evaluate the performance of the proposed model, a simulation study was carried out comparing the proposed

[a]Master's Degree. E-mail: leonardopinheiromiranda@hotmail.com
[b]Ph.D. E-mail: silvana.schneider@ufrgs.br
[c]Ph.D. E-mail: cerqueirajr@ufpa.br

model with models available in the literature (Weibull and piecewise exponential), using the R software. Finally, an application of the proposed model was conducted on a real data set related to personal loans, to evaluate the applicability of the model in a credit context.

***Keywords*:** Credit risk; Long-term survival model; Time-to-default; Weibull.

Resumen

En el contexto financiero, el análisis de supervivencia se ha utilizado de diversas maneras, como en el estudio de Borelli & Lucena (2022), cuyo objetivo es estimar el tiempo hasta la recuperación de las carteras de crédito vencidas, con especial atención a la valoración de las carteras de crédito morosas. Al igual que en el estudio de Ramirez (2016) que buscó estimar el tiempo hasta el incumplimiento, evaluando las variables condicionantes de la demora crediticia. Uno de los objetivos de este trabajo es analizar un conjunto de datos reales sobre crédito para modelar el tiempo hasta que un cliente de una empresa financiera, ubicada en Rio Grande do Sul, Brasil, se convierte en moroso. Por lo tanto, en este trabajo se propone un modelo para datos de supervivencia con tasa de recuperación, en el que la distribución de los tiempos se ajusta mediante la distribución exponencial por tramos de potencia (PPE). El otro objetivo es proponer un modelo aún no explorado conjuntamente en la literatura, mediante la construcción de un modelo de supervivencia a largo plazo, considerando el enfoque de los modelos de tiempo de promoción (Yakovlev & Tsodikov, 1996) y la distribución exponencial por tramos de potencia (Gómez et al., 2017). Para evaluar el rendimiento del modelo propuesto, se realizó un estudio de simulación comparándolo con modelos disponibles en la literatura (Weibull y exponencial por tramos), utilizando el software R. Finalmente, se aplicó el modelo propuesto a un conjunto de datos reales relacionados con préstamos personales para evaluar su aplicabilidad en un contexto crediticio.

***Palabras clave*:** Modelo de supervivencia a largo plazo; Tiempo hasta el impago; Riesgo crediticio; Weibull.

1. Introduction

Numerous studies have examined credit risk modeling, particularly in the context of personal loans. However, the application of survival analysis in this domain remains relatively underexplored. One of the main advantages of survival analysis is the fact that it has mechanisms to deal with censored data. This means that, even when an individual has not yet presented the event of interest, it is possible to use their information in survival analysis, to estimate the probability of default in the future. This feature is especially useful when dealing with credit granting data that have only been active for a short time.

In the financial context, survival analysis has been used in a variety of ways, such as in the study (Borelli & Lucena, 2022), which aims to estimate the time to recovery of overdue credit portfolios, with a focus on pricing nonperforming credit portfolios. As in the study by Ramirez (2016), which aimed to estimate the time

until default, evaluating the covariates associated with the credit delay. One of the objectives of this paper is to analyze a real dataset about credit, to model the time until the customer of a financial company, located in Rio Grande do Sul - Brazil, becomes defaulter.

For this application, we analyzed a database comprising approximately 117 000 unique loan borrowers, covering the period from November 2017 to May 2023. From the perspective of the financial sector, for a company to be profitable, it is quite reasonable to expect that part of the customers who consume credit will always be compliant and consequently "immune" to the event of default. This perspective suggests that the class of cure rate models, as examined in this study, is well suited to this context. The model proposed in this study can serve as a valuable tool for more accurately assessing customer risk. The motivation for evaluating the proposed model in the credit domain is derived from the significance of the sector for the national economy and the growing interest in its recent applications.

There basically are two large classes of cure rate models consolidated in the literature: the first, is called mixture models, which was initially proposed by Boag (1949) and Berkson & Gage (1952), where the latent variable is defined by a Bernoulli distribution; the another class, aiming to consider a competitive risk structure in the modeling, presented by Yakovlev & Tsodikov (1996), is called promotion time, the latent random variable is a Poisson distribution.

In survival analysis with a cure rate, semiparametric models have been widely explored, among which the piecewise exponential model (PEM) stands out as one of the most popular. Proposed by Kalbfleisch & Prentice (1973) and, since then, it has been gaining ground in the literature, given its flexibility. For example, according to some authors, such as Ibrahim et al. (2001), Demarqui (2010), Gómez et al. (2017) and Santos Junior & Schneider (2022), much of its acceptance and popularity stems from the model's flexibility for failure rate functions, which can take different forms.

The use of distributions transformed with the power parameter has been used as an option for existing distributions. An approach proposed by Lehmann (1953) and became known in the literature as the Lehmann alternative model. This approach consists of proposing a probabilistic model based on the power of another continuous and differentiable accumulated distribution. In the context of survival analysis, the use of power distributions has become popular and studies such as Castro & Gómez (2020) and Pal et al. (2022) are beginning to emerge.

In this context, the aim of this paper is to propose a novel model for long-term survival data, combining the promotion time model with the power piecewise exponential distribution (PPE) — a combination not yet explored in the literature. To assess the proposed model, we conducted a simulation study by varying both the data-generating distribution and the number of intervals in the PPE distribution. The objective of the simulation study is to evaluate the performance of the proposed model using the following measures: percentage relative bias (Bias), standard deviation (SD), mean standard error (SE), confidence interval (CI), and coverage probability (PC).

2. Methodology

In the proposal of Yakovlev & Tsodikov (1996) it is assumed that the latent random variable represents the number of competing causes (M), and follow a Poisson distribution with mean θ. The construction of the model can be described as follows.

Let Z_i be a random variable, latent, unobservable variable that denotes the time of promotion of the event of interest due to the i-th cause. Given $M = m$, assume that $Z_i, i = 1, 2, \ldots, m$, independent and identically distributed random variables (i.i.d.), the time until the i-th cause produces the event of interest. Then the cumulative distribution function will be given by $F(z) = 1 - S(z)$. The time for the event of interest to occur will be defined by $T = \min\{Z_i, 0 \leq i \leq M\}$. In this way, the main functions are defined by

$$S_p(t) = e^{-\theta(F_0(t))}; \quad f_p(t) = \theta f_0(t) e^{-\theta(F_0(t)}; \quad h_p(t) = \theta f_0(t), \tag{1}$$

where, $S_p(t)$ is the population survival function; $f_p(t)$ the population probability density function; $h_p(t)$ the population hazard function; $F_0(t)$ represents a cumulative distribution function specific to susceptible individuals; $f_0(t)$ the probability density function for the susceptible individuals; $\theta = \exp(\boldsymbol{x}'\boldsymbol{\beta})$, where $\boldsymbol{x}$ denotes the covariate matrix and $\boldsymbol{\beta}$ denotes the vector of covariates regression.

We can note that $S_p(t)$ will be improper, that is, $\lim_{t\to\infty} S_p(t) = e^{-\theta}$, which represents the cure rate p_o. It is interesting to note that the probability of cure rate, $p_0 = e^{-\theta}$, when $\theta \to \infty$ will converge to $p_o = 0$, that is, absence of cure rate. However, whereas $\theta \to 0$ we will have $p_o = 1$.

The likelihood function take into account the contribution of uncensored individuals through the probability density function, and the contribution of censored observations through the survival function. So, the marginal likelihood function is given by Carneiro (2012), by the following equation:

$$\mathrm{L}(\Theta, p_0) \propto \prod_{i=1}^{n} [f_p(t_i|\Theta, p_0)]^{\delta_i} [S_p(t_i|\Theta, p_0)]^{1-\delta_i}, \tag{2}$$

where Θ is the vector of distribution parameters, δ_i is the censoring indicator, with $\delta_i = 0$ if the individual is censored and $\delta_i = 1$ otherwise and p_0 represents the cure rate.

Because the times are modeled using the power piecewise exponential model (PPEM), an extension of the piecewise exponential model (PEM), we first present the definition of the PEM.

Let $\tau = s_0, s_1, \ldots, s_b$ be the grid that partitions the time axis into b intervals, where $0 = s_0 < s_1 < \cdots < s_b = \infty$. The intervals are defined as $I_j = (s_{j-1}, s_j]$ for $j = 1, 2, \ldots, b$. Within each interval, the time follows an exponential distribution with hazard function λ_j. Consequently, the cumulative hazard function and the survival function are given, respectively, by:

$$H_0(t|\lambda, \tau) = \sum_{j=1}^{b} \lambda_j (t_j - s_{j-1}) \quad ; \quad S_0(t|\lambda, \tau) = \exp\Big\{ - \sum_{j=1}^{b} \lambda_j (t_j - s_{j-1}) \Big\}. \tag{3}$$

Based on the concepts discussed above, we now characterize the proposed model, which is based on the EPP distribution introduced by Gómez et al. (2017), whose cumulative distribution function is given by:

$$\begin{aligned} F(t|\lambda,\tau,\alpha) &= \Big[\{F_0(t|\lambda,\tau)\}\Big]^{\alpha} \\ &= \Big[1-\exp\Big\{-\sum_{j=1}^{b}\lambda_j(t_j-s_{j-1})\Big\}\Big]^{\alpha}, \alpha>0, \end{aligned} \tag{4}$$

where $F_0(t|\lambda,\tau)$ represents the cumulative distribution function of the PE distribution, that is, $F_0(t|\lambda,\tau) = 1-\exp\{-\sum_{j=1}^{b}\lambda_j(t_j-s_{j-1})\}$.

Deriving the cumulative distribution function of $F(t|\lambda,\tau,\alpha)$ in relation to t, we have the density function of the PPE distribution given by

$$\begin{aligned} f(t|\lambda,\tau,\alpha) &= \alpha\Big[F_0(t|\lambda,\tau)\Big]^{\alpha-1} f_0(t|\lambda,\tau) \\ &= \alpha\Big[1-\exp\Big\{-\sum_{j=1}^{b}\lambda_j(t_j-s_{j-1})\Big\}\Big]^{\alpha-1} \\ &\times \lambda_j\Big(1-\Big[1-\exp\Big\{-\sum_{j=1}^{b}\lambda_j(t_j-s_{j-1})\Big\}\Big]\Big). \end{aligned} \tag{5}$$

Given the relation $S(t) = 1-[F(t)]$ and Equation (4), the survival function is given by

$$\begin{aligned} S(t|\lambda,\tau,\alpha) &= 1-[F_0(t|\lambda,\tau)]^{\alpha} \\ &= 1-\Big[1-\exp\Big\{-\sum_{j=1}^{b}\lambda_j(t_j-s_{j-1})\Big\}\Big]^{\alpha}. \end{aligned} \tag{6}$$

To derive the likelihood function of the proposed model, we use the population probability density function given in Equation (5) and the survival function in Equation (6). Accordingly, the log-likelihood function is expressed in Equation (2), resulting in:

$$
\begin{aligned}
\mathrm{l}(\Theta, p_0) &= \sum_{i=1}^{n} \delta_i \log([f_p(t_i|\Theta, p_0)]) + \sum_{i=1}^{n} (1-\delta_i) \log([S_p(t_i|\Theta, p_0)]) \\
&= \sum_{i=1}^{n} \delta_i \log \left[\theta_i \alpha \Big[1 - \exp\Big\{ - \sum_{j=1}^{b} \lambda_j (t_j - s_{j-1}) \Big\} \Big]^{\alpha-1} \right. \\
&\times \lambda_j \Big(1 - \Big[1 - \exp\Big\{ - \sum_{j=1}^{b} \lambda_j (t_j - s_{j-1}) \Big\} \Big] \Big) \\
&\left. \times \exp\Big[- \theta_i \Big(1 - \exp\Big\{ - \sum_{j=1}^{b} \lambda_j (t_j - s_{j-1}) \Big\} \Big)^{\alpha} \Big] \right] \\
&+ \sum_{i=1}^{n} (1-\delta_i) \Big[- \theta_i \Big(1 - \exp\Big\{ - \sum_{j=1}^{b} \lambda_j (t_j - s_{j-1}) \Big\} \Big)^{\alpha} \Big].
\end{aligned}
$$

The estimators are obtained by computing the first-order partial derivatives of the log-likelihood with respect to the parameter vector Θ, yielding the score function $U(\Theta) = \frac{\partial l(\Theta, p_0|t)}{\partial \Theta}$. The maximum likelihood estimator (MLE) corresponds to the solution of the system of equations $U(\hat{\Theta}) = 0$. Furthermore, the second-order derivatives with respect to Θ provide the Hessian matrix, defined as $H(\Theta) = \frac{\partial^2 l(\Theta, p_0|t)}{\partial \Theta^2}$. In this study, the computational estimation was performed using the optim function from the R (R Core Team, 2020), based on the maximum likelihood method.

3. Simulation Study

In order to evaluate the performance of the proposed model, a simulation study was conducted comparing the proposed model with models already established in the literature (Weibull and piecewise exponential). For the simulated scenario, the generation of time follow the distribution PE(λ, τ), with 4 interval, whose parameters were defined with the rate vector $\lambda = \{0.9, 0.7, 0.5, 0.3\}$ and the time grid $\tau = \{0, 0.15, 0.35, 0.85\}$. The regression coefficients were defined as $\beta_0 = -0.6$, $\beta_1 = 0.4$ and $\beta_2 = 0.5$. The data generation process can be summarized, as shown in the Table below.

Data generation
1: Generate uniforms, independent, u_i, $v_i \sim$ Uniforme[0,1];
2: Generate two covariates, $x_{i1} \sim$ Bernoulli(0.5) and $x_{i2} \sim$ Normal$(0, 1)$;
3: $\theta_i = \exp(x_i' \beta)$;
4: If $(u_i <= \exp(-\theta_i)$, assume $t_i = \infty$; otherwise:
4.1: $F^{-1}(u_i) = t_i$;
5: $F^{-1}(v_i) = c_i$;
6: Define $y_i = min\{t_i, c_i\}$;
7: Define the censor as $\delta_i = 1$ when $y_i = t_i$ or $\delta_i = 0$ when $y_i = c_i$.

Across the constructed scenarios, different sample sizes were considered, with $n \in \{250, 500, 1000, 2000\}$ and the cure rate model adjustments were applied, which we denote by

- M0: Model with cure rate and Weibull distribution $Weib(\lambda, \alpha)$;
- M1: Model with cure rate and piecewise exponential distribution $PE(\lambda, \tau)$, with 4 intervals;
- M2: Model with cure rate and power piecewise exponential distribution $PPE(\lambda, \tau, \alpha)$, with 4 intervals.

At the end of each simulation, summary statistics were measured by: percentage relative bias (Bias), standard deviation (SD), mean standard errors (SE), confidence interval (CI), coverage probability (PC), and the criteria AIC and BIC.

In general, can be seen in the Tables 1, 2 and 3 as the sample size increases, for any of the adjustments (M0, M1 and M2), the Bias, SD and SE associated with the estimation of regression coefficients reduce. The M0 adjustment, Table 1, demonstrates that the variation in sample size has less influence on the M0 adjustment, compared to the others adjustments, even for small samples, the relative bias in the estimate of the regression coefficient β_0 does not show great variation. For example, for a sample size of 250, the relative bias of β_0 was 6.191%.

TABLE 1: Adjustment of the model M0 - data generating with the distribution EP{0.9, 0.7, 0.5, 0.3}.

N	Par.	Real	Est.	Bias	SD	SE	CP	CI
250	β_0	-0.600	-0.563	6.191	0.515	0.176	0.998	[-1.573 ; 0.447]
	β_1	0.400	0.391	-2.323	0.193	0.197	0.944	[0.012 ; 0.770]
	β_2	0.500	0.505	0.921	0.103	0.101	0.944	[0.304 ; 0.706]
500	β_0	-0.600	-0.614	-2.250	0.117	0.121	0.958	[-0.843 ; -0.384]
	β_1	0.400	0.410	2.381	0.135	0.139	0.954	[0.145 ; 0.674]
	β_2	0.500	0.504	0.736	0.070	0.072	0.954	[0.366 ; 0.642]
1000	β_0	-0.600	-0.603	-0.430	0.081	0.085	0.950	[-0.761 ; -0.445]
	β_1	0.400	0.404	1.072	0.098	0.098	0.938	[0.212 ; 0.597]
	β_2	0.500	0.505	1.003	0.051	0.050	0.938	[0.405 ; 0.605]
2000	β_0	-0.600	-0.604	-0.654	0.062	0.060	0.948	[-0.725 ; -0.483]
	β_1	0.400	0.408	2.105	0.068	0.069	0.946	[0.276 ; 0.541]
	β_2	0.500	0.499	-0.274	0.033	0.035	0.946	[0.433 ; 0.564]

[1]Par: Parameter; Est: Estimated; Bias: relative bias; SD: standard deviation; SE: mean standard errors; CP: coverage probability; CI: confidence interval.

In adjustment M1, Table 2, it is possible to observe the lack of precision in the estimation of the regression coefficient parameter, β_0 for a small sample (250), in which the parameter is overestimated, presenting a relative bias of 139.814%. However, as the sample size increases, the relative bias value reduces and the estimated parameter approaches the real value. It is also worth highlighting the estimation of λ_j which, despite the bias being greater than the values observed for betas, are still not above 15%.

TABLE 2: Adjustmentof the model M1 - data generating with the distribution EP{0.9, 0.7, 0.5, 0.3}.

N	Par.	Real	Est.	Bias	SD	SE	CP	CI
250	β_0	-0.600	0.239	139.814	1.842	2.255	0.874	[-3.371 ; 3.849]
	β_1	0.400	0.410	2.398	0.219	0.110	0.938	[-0.019 ; 0.838]
	β_2	0.500	0.500	-0.010	0.110	0.056	0.938	[0.284 ; 0.716]
	λ_1	0.900	0.802	-10.940	0.512	2.280	0.982	[-0.201 ; 1.805]
	λ_2	0.700	0.638	-8.827	0.430	2.309	0.982	[-0.204 ; 1.48]
	λ_3	0.500	0.467	-6.553	0.345	2.354	0.974	[-0.209 ; 1.144]
	λ_4	0.300	0.342	13.867	0.308	2.469	0.944	[-0.261 ; 0.945]
500	β_0	-0.600	-0.251	58.243	1.164	0.810	0.946	[-2.532 ; 2.031]
	β_1	0.400	0.413	3.235	0.158	0.077	0.952	[0.104 ; 0.722]
	β_2	0.500	0.508	1.610	0.079	0.039	0.952	[0.353 ; 0.663]
	λ_1	0.900	0.833	-7.481	0.359	0.829	0.912	[0.128 ; 1.537]
	λ_2	0.700	0.671	-4.162	0.313	0.851	0.918	[0.057 ; 1.285]
	λ_3	0.500	0.486	-2.875	0.247	0.886	0.926	[0.002 ; 0.969]
	λ_4	0.300	0.317	5.741	0.194	0.975	0.952	[-0.062 ; 0.697]
1000	β_0	-0.600	-0.563	6.206	0.404	0.156	0.980	[-1.354 ; 0.229]
	β_1	0.400	0.407	1.641	0.111	0.054	0.952	[0.188 ; 0.625]
	β_2	0.500	0.503	0.582	0.053	0.028	0.952	[0.400 ; 0.606]
	λ_1	0.900	0.906	0.645	0.228	0.169	0.948	[0.459 ; 1.352]
	λ_2	0.700	0.705	0.689	0.195	0.183	0.954	[0.322 ; 1.087]
	λ_3	0.500	0.518	3.524	0.161	0.206	0.944	[0.203 ; 0.833]
	λ_4	0.300	0.336	11.960	0.135	0.268	0.944	[0.072 ; 0.600]
2000	β_0	-0.600	-0.584	2.618	0.190	0.081	0.956	[-0.956 ; -0.213]
	β_1	0.400	0.401	0.157	0.076	0.038	0.944	[0.253 ; 0.549]
	β_2	0.500	0.501	0.188	0.035	0.020	0.944	[0.433 ; 0.569]
	λ_1	0.900	0.901	0.119	0.174	0.090	0.944	[0.560 ; 1.242]
	λ_2	0.700	0.708	1.102	0.157	0.100	0.950	[0.401 ; 1.015]
	λ_3	0.500	0.512	2.342	0.127	0.117	0.928	[0.263 ; 0.760]
	λ_4	0.300	0.324	8.101	0.110	0.161	0.928	[0.109 ; 0.540]

[2]Par: Parameter; Est: Estimated; Bias: relative bias; SD: standard deviation; SE: mean standard errors; CP: coverage probability; CI: confidence interval.

Adjustment M2, Table 3 presents results very similar to those observed in the M1, in which for small samples there is an overestimation of the intercept parameter, however, as the sample size increases, we have convergence to the true parameter. Measures of probability of coverage close to nominal value of 0.95 and standard error with low values, for a sample size of 2000, the parameter β_0 presents indexes of 0.078 standard error and 0.960 probability of coverage.

Evaluating the estimates of the regression coefficient β_0 using the Figure **??**,can be seen that both the M1 and M2 adjustments presented greater variations in the estimation of the parameter β_0 for small samples, but as the sample size increases, they become closer to the true value. It can also be observed the overestimation of the β_0 parameter, when the adjustment time grid is not the same as the data generation specification.

TABLE 3: Adjustment of the model M2 - data generating with the distribution EP{0.9, 0.7, 0.5, 0.3}.

N	Par.	Real	Est.	Bias	SD	SE	CP	CI
250	β_0	-0.600	0.290	148.349	2.067	3.169	0.882	[-3.762 ; 4.342]
	β_1	0.400	0.410	2.455	0.219	0.110	0.940	[-0.019 ; 0.839]
	β_2	0.500	0.500	0.043	0.110	0.056	0.940	[0.284 ; 0.717]
	λ_1	0.900	1.125	25.033	1.025	3.310	0.946	[-0.883 ; 3.134]
	λ_2	0.700	0.718	2.570	0.502	3.262	0.964	[-0.266 ; 1.702]
	λ_3	0.500	0.502	0.486	0.368	3.288	0.966	[-0.219 ; 1.224]
	λ_4	0.300	0.360	19.874	0.318	3.393	0.942	[-0.264 ; 0.984]
	α_p	1.000	1.128	12.844	0.305	0.129	0.934	[0.532 ; 1.725]
500	β_0	-0.600	-0.203	66.170	1.396	1.331	0.946	[-2.939 ; 2.533]
	β_1	0.400	0.413	3.256	0.158	0.077	0.950	[0.104 ; 0.722]
	β_2	0.500	0.508	1.634	0.079	0.039	0.950	[0.353 ; 0.664]
	λ_1	0.900	0.974	8.230	0.557	1.461	0.958	[-0.118 ; 2.066]
	λ_2	0.700	0.712	1.655	0.348	1.420	0.920	[0.030 ; 1.394]
	λ_3	0.500	0.505	1.034	0.258	1.439	0.972	[0.000 ; 1.011]
	λ_4	0.300	0.327	8.942	0.196	1.519	0.950	[-0.058 ; 0.711]
	α_p	1.000	1.055	5.525	0.169	0.084	0.936	[0.725 ; 1.386]
1000	β_0	-0.600	-0.555	7.477	0.454	0.188	0.972	[-1.444 ; 0.334]
	β_1	0.400	0.407	1.642	0.112	0.054	0.952	[0.188 ; 0.625]
	β_2	0.500	0.503	0.587	0.053	0.028	0.952	[0.400 ; 0.606]
	λ_1	0.900	0.955	6.116	0.363	0.278	0.954	[0.243 ; 1.667]
	λ_2	0.700	0.715	2.120	0.228	0.244	0.940	[0.268 ; 1.162]
	λ_3	0.500	0.521	4.210	0.174	0.256	0.944	[0.179 ; 0.863]
	λ_4	0.300	0.338	12.521	0.140	0.313	0.950	[0.064 ; 0.612]
	α_p	1.000	1.018	1.776	0.111	0.058	0.948	[0.801 ; 1.235]
2000	β_0	-0.600	-0.597	0.483	0.168	0.078	0.960	[-0.926 ; -0.269]
	β_1	0.400	0.401	0.350	0.076	0.038	0.948	[0.253 ; 0.550]
	β_2	0.500	0.501	0.146	0.039	0.020	0.948	[0.424 ; 0.578]
	λ_1	0.900	0.927	2.966	0.260	0.139	0.948	[0.418 ; 1.436]
	λ_2	0.700	0.715	2.118	0.166	0.114	0.952	[0.390 ; 1.040]
	λ_3	0.500	0.519	3.867	0.129	0.122	0.938	[0.267 ; 0.772]
	λ_4	0.300	0.327	8.853	0.104	0.162	0.942	[0.123 ; 0.531]
	α_p	1.000	1.004	0.359	0.083	0.040	0.954	[0.840 ; 1.167]

[3]Par: Parameter; Est: Estimated; Bias: relative bias; SD: standard deviation; SE: mean standard errors; CP: coverage probability; CI: confidence interval.

In addition, to evaluating the measures previously observed, such as relative bias, coverage probability, mean standard errors and standard deviation, the AIC and BIC criteria. This indication is due to the lowest observed value of the AIC and BIC adjustment criteria among all the models tested.

Through the simulation study carried out, it was possible to see that the proposed adjustment (M2) presented estimation quality better than the models known in the literature. Remembering that the data generating model is M1, which is equivalent to the PPE model with power parameter $\alpha = 1$. Where the values observed for the measures of relative bias, standard deviation, mean standard errors and coverage probability had slightly better accuracy than the M1 model, as can be seen in the Tables 3 and 2.

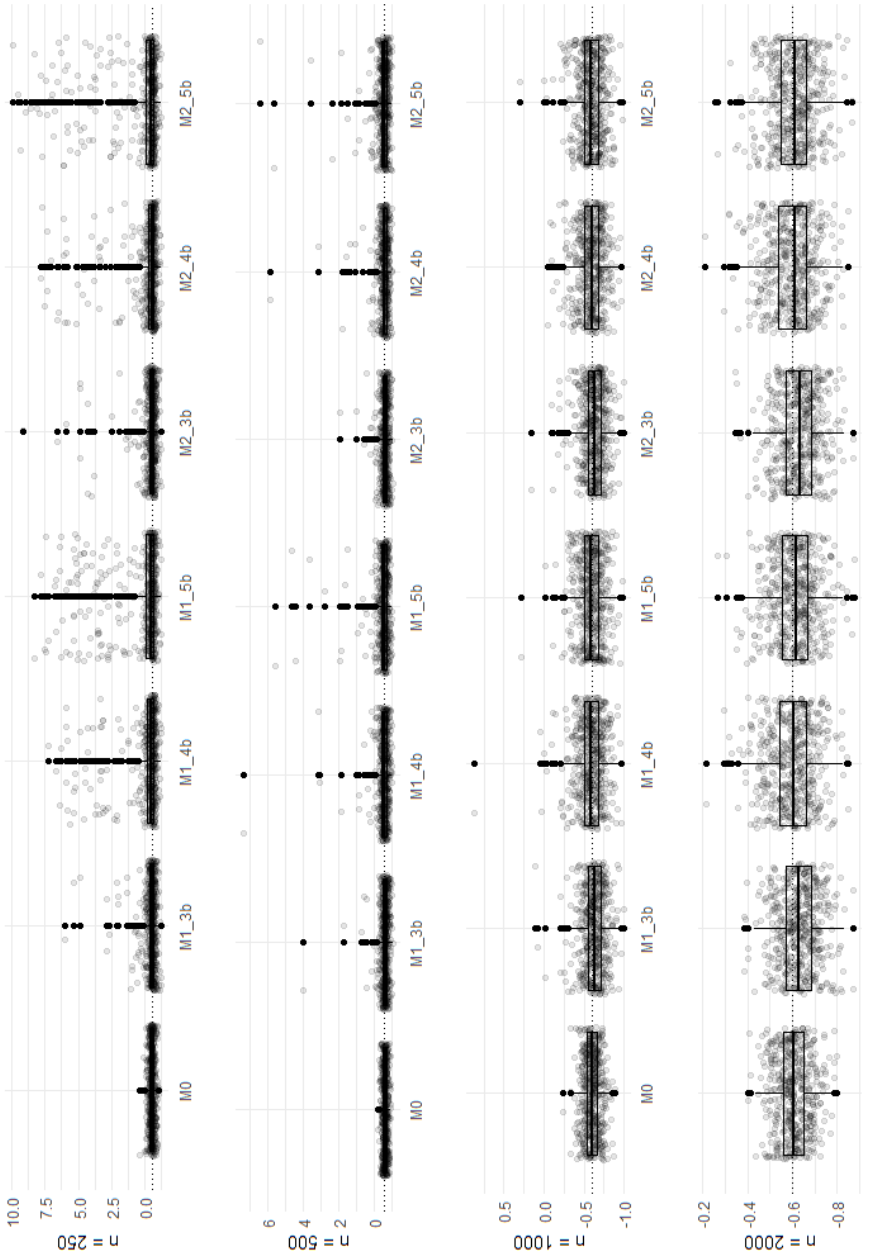


FIGURE 1: Comparison of the estimated β_0 of all adjusted models, considering different sample sizes

4. Credit Scoring Data Analysis

In the business sector of the credit, one of the pillars of sustainability and corporate prosperity consists of the management and correct measurement of credit risk (González-Hermosillo, 1999). According to Caouette et al. (1998), the credit risk represents the risk associated with the non-compliance of the payment obligations, signed by the credit borrower, in full or within the established term. The effects of default can generate negative consequences for all parties involved, that is, both for the credit borrower, as well as for creditor companies and the economy in general (Saeed & Izzeldin, 2016). Therefore, default is a key part of any company.

Therefore, statistical techniques for analyzing credit data have been used in order to allow a more precise assessment of the customer profile, as well as the estimation of the associated risk of future loss. Among the techniques widely accepted in the credit context, due to the ease of implementation and interpretation of results, are the following: logistic regression, as seen in the works of Albuquerque et al. (2017) and Araújo & de Montreuil Carmona (2007); the discriminant analysis used in Araújo & Carmona (2009) and Silva (2017) and decision trees present in the papers of Prazeres Filho (2014) and Moraes Sousa & Figueiredo (2014). In the credit context, survival analysis has been used in very different ways, as we can see in the study Borelli & Lucena (2022) and Ramirez (2016).

4.1. Time Until Default - Credit Loan

This paper proposes to estimate the time until the customer of a financial institution in RS defaults. Thus a sample of customers who took out a personal loan was considered, from November 2017 to May 2023. The final data set has 82,033 customers, containing the following covariates: income, age, region concession, concession risk (credit score), professional occupation, gender and delay flag (incidence or not of delay greater than or equal to 90 days).

In the literature review there is no unanimity for the definition of default, in the theoretical reference, leaving the definition of the concept open to discussion, respecting the particularities of each business. Therefore, for this paper, the concept of problematic assets specified in resolution 4557, which determines default as credits overdue for more than ninety days, by the sector's regulatory, the Central Bank of Brazil was choosen (Banco Central do Brasil, 2006).

4.2. Descriptive Analysis

The Table 4 shows the profile of customers studied.

- The income distribution of customers who acquired a loan with the institution studied presented an average of $2,967$ in reais and a standard deviation of similar magnitude $3,057$ in reais;
- The age distribution of customers has an average of 40.51 years.

- Most of the clients studied, around two-thirds of the base, are classified as low credit risk. It can be observed that the proportion of customers who fail is higher among those classified as high risk, whereas customers classified as low risk exhibit a lower failure proportion. This indicates that the company's risk classification effectively discriminates between low and high risk customers.
- About the retail company's operating niche, as seen in the macro economic indicators, we have that the Northeast and Web regions (proposals approved directly in the company's application, focusing on the digital public) presented higher proportions of customers with a delay equal to or greater than 90 days (failure). On the other hand the Central West region with the lowest default rates.
- Assessing the professional occupation profile of the base, the large part of the clients studied are concentrated in the "Salaried" and "Self-employed/Freelance/Owner" classes, accounting in total for around 85% of the total number of clients.

Additionally to the descriptive analysis, the Kaplan-Meier estimator was performed, it can be seen that the survival curve presents a plateau, indicating the presence of a cure rate. According to Yu et al. (2004), a way to evaluate the consistency of the observed cure rate is to compare the distance between the median survival time, of those who presented the event of interest, with the data follow-up time. If the follow-up time is relatively longer than the median of those who presented the event of interest, we have an indication that the cure rate actually exists. In Figure 2, it can be seen that the median of the failure time is close to 5 months, while the follow-up time is greater than 60 months. So, it reinforces the presence of an observed cure rate.

Table 4: Table with the summary statistics.

		Total		Failure		Censoring	
Covariates		Mean	(SD)	Mean	(SD)	Mean	(SD)
Income		2.967.17	(3057.22)	2.209.98	(2076.06)	3.406.31	(1077.02)
Age		40.51	(13.86)	38.25	(12.7)	41.82	(9.75)
	Categories	Freq.	(%)	Freq.	(%)	Freq.	(%)
Credit Risk	Low	61.743	(75.27)	19.609	(65.12)	42.134	(81.15)
	High	20.290	(24.73)	10.503	(34.88)	9.787	(18.85)
Gender	F	50.883	(62.03)	16.549	(54.96)	34.334	(66.13)
	M	31.150	(37.97)	13.563	(45.04)	17.587	(33.87)
Region	R1 (S e CO)	15.661	(19.09)	5.055	(16.79)	10.606	(20.43)
	R2 (SE e NO)	37.043	(45.16)	13.029	(43.27)	24.014	(46.25)
	R3 (NE e WEB)	29.329	(35.75)	12.028	(39.94)	17.301	(33.32)
Ocupacion	Retiree	11.516	(14.04)	2.975	(9.88)	8.541	(16.45)
	Salaried	37.970	(46.29)	13.886	(46.11)	24.084	(46.39)
	Self-employed	32.547	(39.67)	13.251	(44)	19.296	(37.16)

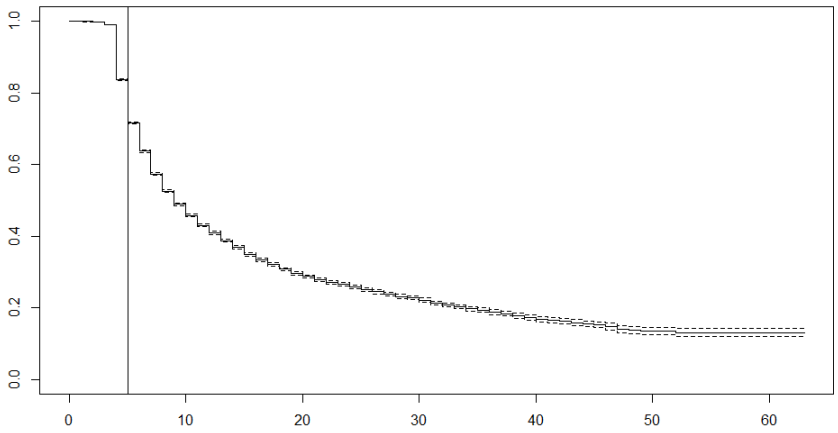


FIGURE 2: Comparison of follow-up time and median of the failure time (vertical line).

4.3. Results - Adjustments

In order to make a good inference, nine adjustments were performed, denote by:

- M0: Model with cure rate and Weibull distribution $Weib(\lambda, \alpha)$;
- M1.2b: Model with cure rate and piecewise exponential distribution $PE(\lambda, \tau)$, with 2 intervals;
- M1.3b: Model with cure cure rate and piecewise exponential distribution $PE(\lambda, \tau)$, with 3 intervals;
- M1.4b: Model with cure cure rate and piecewise exponential distribution $PE(\lambda, \tau)$, with 4 intervals;
- M1.5b: Model with cure cure rate and piecewise exponential distribution $PE(\lambda, \tau)$, with 5 intervals;
- M2.2b: Model with cure rate and power piecewise exponential distribution $PPE(\lambda, \tau, \alpha)$, with 2 intervals.
- M2.3b: Model with cure rate and power piecewise exponential distribution $PPE(\lambda, \tau, \alpha)$, with 3 intervals;
- M2.4b: Model with cure rate and power piecewise exponential distribution $PPE(\lambda, \tau, \alpha)$, with 4 intervals;

- M2.5b: Model with cure rate and power piecewise exponential distribution $PPE(\lambda, \tau, \alpha)$, with 5 intervals.

Based on the information criteria AIC and BIC, among the models applied observing Table 5, we have that the M2 adjustment with 2 intervals is the one with the lowest AIC. However, the estimated cure rate is far from the expected value, through Kaplan-Meier (0.1512). Taking into account both the AIC criteria and the cure rate estimate, the M2.3b adjustment is the best choice and therefore this adjustment is presented in Table 6, without the reference categories. To assess

Table 5: AIC and BIC criteria for selecting the best adjustment.

Adjustment model	AIC	BIC	Cure rate
M0	210.009	210.112	0.4065
M1.2b	163.895	164.009	0.0244
M1.3b	180.613	180.736	0.1343
M1.4b	181.885	182.019	0.0376
M1.5b	247.137	247.281	0.1023
M2.2b	157.163	157.287	0.0215
M2.3b	166.967	167.101	0.1282
M2.4b	157.570	157.714	0.0332
M2.5b	217.024	217.179	0.0954

the significance of the covariates, the "wald.test" function, available in software (R Core Team, 2020) was used. Obtained a significant p-value for all covariates, considering a significance level of α of 0.05.

Therefore, in the model used for the final adjustment, a power piecewise exponential (PPE) distribution with 3 intervals was considered. The covariates were considered in the cure fraction parameter, as follows

$$h_p(t) = \theta f_0(t), \ \theta = \exp(\boldsymbol{x}'\boldsymbol{\beta}),$$

where $\boldsymbol{x}$ denotes the covariate matrix with X = (Income, Age, Gender, Credit risk, Ocupacion.

To interpret the regression coefficients estimated by the adjustments, the sign of the estimated value will provide information on whether the presence of that variable contributes as a protective factor or over-risk factor to the risk of default over time. If the sign of the coefficient is negative, it means that the variable contributes to reducing the risk of default, while if it is positive it increases the risk of default.

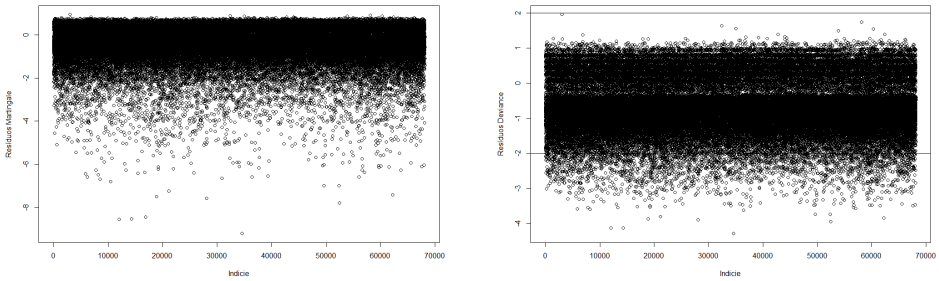
The estimated values of the regression coefficients of the adjustments M0, M1 and M2 presented similar values between the adjustments studied. Therefore, the interpretation of the influence of covariates on survival time for all settings studied was similar. Among the variables studied, Income and Age contributed as protective factor to the risk of default, that is, as the customer's income or age increases, reduces the risk associated with default. Fora all others variables, the presence of categorical variables in relation to the reference categorical contributed to an increase in the risk of default, see Table 6.

TABLE 6: Estimated coefficients.

	Est.	exp(Est.)	SE	CI
M0		**AIC 210.009**	**BIC 210.112**	**Cure rate 0.407**
β_0	-0.105	0.900	0.034	[-0.171 ; -0.039]
β_{Income}	-0.403	0.669	0.010	[-0.423 ; -0.382]
β_{Age}	-0.111	0.895	0.007	[-0.125 ; -0.098]
β_{Male}	0.371	1.449	0.012	[0.348 ; 0.393]
β_{Reg2}	0.149	1.161	0.028	[0.094 ; 0.204]
β_{Reg3}	0.211	1.235	0.028	[0.155 ; 0.267]
$\beta_{Highrisk}$	0.349	1.418	0.012	[0.325 ; 0.373]
$\beta_{Salaried}$	0.110	1.117	0.022	[0.067 ; 0.154]
$\beta_{Selfemployed}$	0.288	1.334	0.105	[0.494 ; 0.082]
M1.3b		**AIC 180.613**	**BIC 180.736**	**Cure rate 0.134**
β_0	0.697	2.007	0.019	[0.659 ; 0.734]
β_{Income}	-0.459	0.632	0.006	[-0.471 ; -0.447]
β_{Age}	-0.132	0.876	0.004	[-0.14 ; -0.125]
β_{Male}	0.418	1.519	0.007	[0.405 ; 0.431]
β_{Reg2}	0.184	1.202	0.016	[0.152 ; 0.215]
β_{Reg3}	0.366	1.442	0.016	[0.334 ; 0.398]
$\beta_{Highrisk}$	0.385	1.470	0.007	[0.371 ; 0.399]
$\beta_{Salaried}$	0.175	1.191	0.013	[0.150 ; 0.200]
$\beta_{Selfemployed}$	0.381	1.464	0.013	[0.356 ; 0.406]
M2.3b		**AIC 166.967**	**BIC 167.101**	**Cure rate 0.128**
β_0	0.720	2.054	0.0192	[0.682 ; 0.757]
β_{Income}	-0.463	0.629	0.0061	[-0.475 ; -0.451]
β_{Age}	-0.134	0.875	0.0039	[-0.142 ; -0.126]
β_{Male}	0.422	1.525	0.0067	[0.409 ; 0.435]
β_{Reg2}	0.187	1.206	0.0162	[0.155 ; 0.219]
β_{Reg3}	0.375	1.455	0.0165	[0.343 ; 0.407]
$\beta_{Highrisk}$	0.387	1.473	0.0071	[0.373 ; 0.401]
$\beta_{Salaried}$	0.178	1.195	0.0128	[0.153 ; 0.203]
$\beta_{Selfemployed}$	0.387	1.473	0.0127	[0.362 ; 0.412]
λ_1	0.003	1.003	0.0001	[0.003 ; 0.003]
λ_2	0.854	2.348	0.0079	[0.838 ; 0.869]
λ_3	11.603	109.464	0.0058	[11.592 ; 11.615]
α_p	0.950	2.586	0.0001	[0.950 ; 0.950]

Furthermore, to evaluate the quality of the model adjustment, the use of graphical analysis of residuals and the estimated survival curve of the adjusted model can help in interpreting the results. In the present work, the martingale and deviance residuals will be evaluated, as well as the comparison of the estimated survival curve of the adjusted model $M2.3b$ versus the Kaplan-Meier adjustment.

Evaluating the martingale and deviance residuals, from the adjustment of the proposed $M2.3b$ model, presented in Figure 3, it is possible to observe the existence of some points whose residuals exceed the value of -2 on deviance. However, it is important to consider the frequency of these observations to the size of the data set studied. Given that the study's analysis set has almost 70 thousand observations, the proportion of outlier points represents around 3.8% of the total observations.



(a) Residual Martingale (b) Residual Deviance

FIGURE 3: Residual analysis – Adjusted model M2.3b.

Considering both the application results and the model selection criteria, the M2 adjustment proves to be a promising approach for modeling time to default, as it additionally allows for the estimation of the default cure rate. Moreover, it enables the estimation of the covariates' effects on survival time.

5. Conclusion

The objective of this paper is to propose a novel combination not yet explored in the literature: a cure rate model based on the promotion time approach with a Poisson latent distribution and a power piecewise exponential distribution for time adjustment (M2), developed within a frequentist framework and applied to the credit context. After characterizing the proposed model, a simulation study was conducted to compare it with two established models from the literature (M1 and M0), with the aim of evaluating the consistency and efficiency of the estimators.

The simulation study shows that the proposed model exhibits relative bias, standard deviation, mean standard errors, and coverage probabilities comparable to those of established models (M0 and M1), and in some cases, much better. As the sample size increased, a reduction in relative bias was observed, indicating that the estimator converges to the true parameter and, therefore, is consistent. A similar pattern was observed for the standard deviation, with larger sample sizes leading to a decrease in the estimator's variance.

The results from the simulated scenarios suggest opportunities to further explore the proposed model by varying the parameters of the time grid and failure rates to improve its performance. As future work, the scenarios developed in this study could be extended by altering the data-generating distribution, adjusting the failure rate grid parameters, incorporating a frailty term, applying zero-inflated models, or adopting methodologies to determine the optimal time grid. Several authors, such as Demarqui (2010), highlight that one of the main challenges in using the PE distribution is determining the appropriate number of intervals.

In addition, the proposed model was applied to analyze loan default data, aiming to model the time until a customer of a financial institution from Rio

Grande do Sul, Brazil, becomes default (defined as a delay of more than 90 days). Consistent with the simulation results, the Application showed similar parameter estimates across the models tested. Based on the AIC selection criterion, the proposed model with three intervals (the cure rate model with a power piecewise exponential distribution $EPP(\lambda, \tau, \alpha)$ yielded the lowest AIC value, indicating the best fit among the models considered. This outcome supports the suitability of the proposed model for this context. For future work, it would be valuable to explore applications focusing on credit recovery, given the relevance of such studies to the financial sector and the potential for this methodology to provide meaningful contributions to the corporate environment.

[Received: September 2025 — Accepted: November 2025]

References

Albuquerque, P. H. M., Medina, F. A. S. & da Silva, A. R. (2017), 'Regressão logística geograficamente ponderada aplicada a modelos de credit scoring', *Revista Contabilidade & Finanças* **28**(73), 93–112.

Araújo, E. A. & Carmona, C. U. D. M. (2009), 'Construção de modelos credit scoring com análise discriminante e regressão logística para a gestão do risco de inadimplência de uma instituição de microcrédito', *Revista eletrônica de administração* **15**(1), 50–77.

Araújo, E. A. & de Montreuil Carmona, C. U. (2007), 'Desenvolvimento de modelos credit scoring com abordagem de regressão logística para a gestão da inadimplência de uma instituição de microcrédito', *Contabilidade Vista & Revista* **18**(3), 107–131.

Banco Central do Brasil (2006), Core principles for effective banking supervision, Technical report, Banco Central do Brasil. Accessed 18 August 2023. `https://www.bcb.gov.br/fis/supervisao/docs/core_principles_traducao2006.pdf`

Berkson, J. & Gage, R. P. (1952), 'Survival curve for cancer patients following treatment', *Journal of the American Statistical Association* **47**(259), 501–515.

Boag, J. W. (1949), 'Maximum likelihood estimates of the proportion of patients cured by cancer therapy', *Journal of the Royal Statistical Society. Series B (Methodological)* **11**(1), 15–53.

Borelli, E. & Lucena, B. D. (2022), 'Recuperação de crédito estimada pela análise de sobrevivência: Credit recovery estimated by survival analysis', *Revista de Gestão e Secretariado (Management and Administrative Professional Review)* **13**(3), 949–973.

Caouette, J. B., Altman, E. I. & Narayanan, P. (1998), *Managing credit risk: the next great financial challenge*, Frontiers in Finance Series, Editora Wiley.

Carneiro, H. P. d. A. (2012), Testes de hipóteses em modelos de sobrevivência com fração de cura, Mestrado em estatística, Universidade Federal do Rio Grande do Norte, Rio Grande do Norte.

Castro, M. & Gómez, Y. M. (2020), 'A bayesian cure rate model based on the power piecewise exponential distribution', *Methodology and Computing in Applied Probability* **22**(2), 677–692.

Demarqui, F. N. (2010), Uma classe mais flexível de modelos semiparamétricos para dados de sobrevivência, Doutorado em estatística, Universidade Federal de Minas Gerais, Minas Gerais.

Gómez, Y. M., Gallardo, D. I. & Arnold, B. C. (2017), 'The power piecewise exponential model', *Journal of Statistical Computation and Simulation* **88**(5), 825–840.

González-Hermosillo, M. B. (1999), *Determinants of ex-ante banking system distress: A macro-micro empirical exploration of some recent episodes*, Editora International Monetary Fund.

Ibrahim, J. G., Chen, M.-H. & Sinha, D. (2001), 'Bayesian semiparametric models for survival data with a cure fraction', *Biometrics* **57**(2), 383–388.

Kalbfleisch, J. D. & Prentice, R. L. (1973), 'Marginal likelihoods based on cox's regression and life model', *Biometrika* **60**(2), 267–278.

Lehmann, E. L. (1953), 'The power of rank tests', *The Annals of Mathematical Statistics* **24**(1), 23 – 43.

Moraes Sousa, M. & Figueiredo, R. S. (2014), 'Análise de crédito por meio de mineração de dados: aplicação em cooperativa de crédito', *JISTEM-Journal of Information Systems and Technology Management* **11**(2), 379–396.

Pal, S., Barui, S., Davies, K. & Mishra, N. (2022), 'A stochastic version of the em algorithm for mixture cure model with exponentiated weibull family of lifetimes', *Journal of Statistical Theory and Practice* **16**(3), 48.

Prazeres Filho, J. (2014), Capacidade preditiva de modelos credit scoring em inferência dos rejeitados, Mestrado em estatística, Universidade Federal de São Carlos, São Paulo.

R Core Team (2020), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ramirez, D. B. (2016), Com uma lupa e um binóculo: uma investigação sobre a inadimplência de um banco de desenvolvimento através de análise de sobrevivência, Mestrado em economia, Universidade Federal de Santa Catarina, Santa Catarina.

Saeed, M. & Izzeldin, M. (2016), 'Examining the relationship between default risk and efficiency in islamic and conventional banks', *Journal of Economic Behavior & Organization* **132**(1), 127–154.

Santos Junior, P. C. & Schneider, S. (2022), 'Power piecewise exponential model for interval-censored data', *Journal of Statistical Theory and Practice* **16**(2), 26.

Silva, P. C. (2017), Regressão logística e análise discriminante na predição da recuperação de portfólios de créditos do tipo non-performing loans, Mestrado em engenharia de produção, Universidade Nove de Julho, São Paulo.

Yakovlev, A. & Tsodikov, A. (1996), *Stochastic models of tumor latency and their biostatistical applications*, Vol. 1, Editora World Scientific, Editora Singapore.

Yu, B., Tiwari, R. C., Cronin, K. A. & Feuer, E. J. (2004), 'Cure fraction estimation from the mixture cure models for grouped survival data', *Statistics in medicine* **23**(11), 1733–1747.